Claims

We claim:

1. A method for approximating a number of tuples returned by a database query that comprises a set of predicates that each reference a set of database tables, the method comprising the steps of:

a) expressing the query as a query selectivity;

b) determining if the query is separable and if so separating the query selectivity to form a product of query selectivity factors;

c) if the query is not separable, decomposing the query selectivity to form a product that comprises a conditional selectivity expression;

d) recursively performing steps b) - f) to determine a selectivity value for each query selectivity factor;

e) matching any conditional selectivity expression with stored statistics to obtain statistics that can estimate the selectivity of the conditional selectivity expressions and using the statistics to obtain an estimated selectivity value; and

f) combining the selectivity values obtained in step d) and the estimated selectivity values obtained in step e) to estimate the selectivity of the query.

2. The method of claim 1 comprising the step of multiplying the estimated selectivity by a Cartesian product of the tables referenced by the predicates to obtain a cardinality of the query.

3. The method of claim 1 wherein the step of separating the query selectivity is performed by separating the predicates that reference different sets of database tables to form a product of query selectivity factors that reference different sets of database tables.

4. The method of claim 1 wherein the product formed in step c) further comprises a query selectivity factor and wherein steps b) - f) are recursively performed to determine a selectivity value for the query selectivity factor in step c).

5. The method of claim 1 wherein steps b) - f) are recursively performed until a non-separable query selectivity that can only be decomposed into a single conditional selectivity expression results.

6. The method of claim 1 comprising the step of storing the estimated selectivity of the query obtained in step f) in memory.

7. The method of claim 6 comprising the step of first determining whether an estimated selectivity is stored for a query and returning that value to approximate the number of tuples returned by the query.

8. The method of claim 1 comprising the step of associating an error with the estimated selectivity value that is based on an accuracy with which the statistic matched

with the conditional selectivity expression can estimate the selectivity of the conditional selectivity expression.

9. The method of claim 8 comprising the step of combining the error associated with each conditional selectivity expression to obtain an estimated error for the selectivity estimation for the query.

10. The method of claim 1 wherein the stored statistics comprise histograms on results of previously executed query expressions.

11. The method of claim 1 wherein the step of matching the conditional selectivity expressions with stored statistics is performed by:

compiling a set of candidate statistics that can be used to estimate the selectivity of the conditional selectivity expression; and

selecting candidate statistics to estimate the selectivity of the conditional selectivity expression based on a selection criteria.

12. The method of claim 11 wherein the selection criteria for a candidate statistic is determined by computing a number of independence assumptions that are made when the candidate is used to estimate the selectivity of the conditional selectivity expression and the selection criteria is to select the candidate that results in the least number of independence assumptions.

13. The method of claim 11 wherein the selection criteria for a candidate statistic is determined by comparing the candidate statistic with a base statistic over the

same column as the candidate statistic and assigning a difference value to the candidate statistic based on a level of difference between the candidate statistic and the base statistic.

14. The method of claim 11 wherein the step of compiling a set of candidate statistics is performed by including statistics that are on results of queries having the same tables referenced by the conditional selectivity expression or a subset of the tables referenced by the conditional selectivity expression and the same predicates over the tables referenced in the conditional selectivity expression or a subset of the predicates over the tables referenced in the conditional selectivity expressions.

15. The method of claim 1 wherein the steps of decomposing the query selectivity and matching the conditional selectivity expressions are repeated to generate alternative products and wherein one of those products is selected to estimate the selectivity of the query.

16. The method of claim 15 wherein the step of decomposing the query is done by exhausting every alternative way of decomposing the query.

17. The method of claim 15 wherein the steps of decomposing the query selectivity to form products of conditional selectivity expressions is performed based on an optimizer search strategy.

18. The method of claim 1 wherein the query is disjunctive and comprising the step of transforming the disjunctive predicates into conjunctive predicates by performing a De Morgan transformation on the disjunctive query.

19. The method of claim 1 wherein the query comprises a GROUP BY predicate over a grouping column and wherein the query is transformed prior to performance of the method steps to return a number of distinct values in the grouping column.

20. The method of claim 19 wherein the step of decomposing the query selectivities performed by considering decompositions that are induced by coalescing grouping.

21. The method of claim 19 wherein the step of decomposing the query selectivities performed by considering decompositions that are induced by invariant grouping.

22. The method of claim 1 wherein the stored statistics comprise histograms built over computed columns in a query result.

23. For use with a database system, a computer readable medium having computer executable instructions stored thereon for performing method steps to approximate a number of tuples returned by a database query that comprises a set of predicates that each reference a set of database tables, the method comprising the steps of:

a) expressing the query as a query selectivity;

b) determining if the query is separable and if so separating the query selectivity to form a product of query selectivity factors;

c) if the query is not separable, decomposing the query selectivity to form a product that comprises a conditional selectivity expression;

d) recursively performing steps b) - f) to determine a selectivity value for each query selectivity factor;

e) matching any conditional selectivity expression with stored statistics to obtain statistics that can estimate the selectivity of the conditional selectivity expressions and using the statistics to obtain an estimated selectivity value; and

f) combining the selectivity values obtained in step d) and the estimated selectivity values obtained in step e) to estimate the selectivity of the query.

24. The computer readable medium of claim 23 comprising the step of multiplying the estimated selectivity by a Cartesian product of the tables referenced by the predicates to obtain a cardinality of the query.

25. The computer readable medium of claim 23 wherein the step of separating the query selectivity is performed by separating the predicates that reference different sets of database tables to form a product of query selectivity factors that reference different sets of database tables.

26. The computer readable medium of claim 23 wherein the product formed in step c) further comprises a query selectivity factor and wherein steps b) - f) are recursively performed to determine a selectivity value for the query selectivity factor in step c).

27. The computer readable medium of claim 23 wherein steps b) - f) are recursively performed until a non-separable query selectivity that can only be decomposed into a single conditional selectivity expression results.

28. The computer readable medium of claim 23 comprising the step of storing the estimated selectivity of the query obtained in step f) in memory.

29. The computer readable medium of claim 28 comprising the step of first determining whether an estimated selectivity is stored for a query and returning that value to approximate the number of tuples returned by the query.

30. The computer readable medium of claim 23 comprising the step of associating an error with the estimated selectivity value that is based on an accuracy with which the statistic matched with the conditional selectivity expression can estimate the selectivity of the conditional selectivity expression.

31. The computer readable medium of claim 30 comprising the step of combining the error associated with each conditional selectivity expression to obtain an estimated error for the selectivity estimation for the query.

32. The computer readable medium of claim 23 wherein the stored statistics comprise histograms on results of previously executed query expressions.

33. The computer readable medium of claim 23 wherein the step of matching the conditional selectivity expressions with stored statistics is performed by:

compiling a set of candidate statistics that can be used to estimate the selectivity of the conditional selectivity expression; and

selecting candidate statistics to estimate the selectivity of the conditional selectivity expression based on a selection criteria.

34. The computer readable medium of claim 33 wherein the selection criteria for a candidate statistic is determined by computing a number of independence assumptions that are made when the candidate is used to estimate the selectivity of the conditional selectivity expression and the selection criteria is to select the candidate that results in the least number of independence assumptions.

35. The computer readable medium of claim 33 wherein the selection criteria for a candidate statistic is determined by comparing the candidate statistic with a base statistic over the same column as the candidate statistic and assigning a difference value to the candidate statistic based on a level of difference between the candidate statistic and the base statistic.

36. The computer readable medium of claim 33 wherein the step of compiling a set of candidate statistics is performed by including statistics that are on results of queries having the same tables referenced by the conditional selectivity expression or a subset of the tables referenced by the conditional selectivity expression and the same predicates over the tables referenced in the conditional selectivity expression or a subset of the predicates over the tables referenced in the conditional selectivity expressions.

37. The computer readable medium of claim 23 wherein the steps of decomposing the query selectivity and matching the conditional selectivity expressions are repeated to generate alternative products and wherein one of those products is selected to estimate the selectivity of the query.

38. The computer readable medium of claim 37 wherein the step of decomposing the query is done by exhausting every alternative way of decomposing the query.

39. The computer readable medium of claim 37 wherein the steps of decomposing the query selectivity to form products of conditional selectivity expressions is performed based on an optimizer search strategy.

40. The computer readable medium of claim 23 wherein the query is disjunctive and comprising the step of transforming the disjunctive predicates into conjunctive predicates by performing a De Morgan transformation on the disjunctive query.

41. The computer readable medium of claim 23 wherein the query comprises a GROUP BY predicate over a grouping column and wherein the query is transformed prior to performance of the method steps to return a number of distinct values in the grouping column.

42. The computer readable medium of claim 41 wherein the step of decomposing the query selectivities performed by considering decompositions that are induced by coalescing grouping.

43. The computer readable medium of claim 41 wherein the step of decomposing the query selectivities performed by considering decompositions that are induced by invariant grouping.

44. The computer readable medium of claim 23 wherein the stored statistics comprise histograms built over computed columns in a query result.

45. An apparatus for approximating a number of tuples returned by a database query that comprises a set of predicates that each reference a set of database tables comprising:

a) means for expressing the query as a query selectivity;

b) means for determining if the query is separable;

c) means for separating the query selectivity to form a product of query selectivity factors if the query is separable;

c) means for decomposing the query selectivity to form a product that comprises a conditional selectivity expression if the query is not separable;

d) means for recursively performing steps b) - f) to determine a selectivity value for each query selectivity factor;

e) means for matching any conditional selectivity expression with stored statistics to obtain statistics that can estimate the selectivity of the conditional selectivity expressions and means for using the statistics to obtain an estimated selectivity value; and

f) means for combining the selectivity values obtained in step d) and the estimated selectivity values obtained in step e) to estimate the selectivity of the query.

46. A method for approximating a number of tuples returned by a database query that comprises a set of predicates that each reference a set of database tables, the method comprising the steps of:

a) expressing the query as a query selectivity;

b) determining if the query is separable and if so separating the query selectivity by separating the predicates that reference different sets of database tables to form a product of query selectivity factors that reference different sets of database tables;

c) if the query is not separable, repeatedly decomposing the query selectivity to form a product that comprises a conditional selectivity expression to generate alternative products and wherein one of those products is selected to estimate the selectivity of the query;

d) recursively performing steps b) - f) to determine a selectivity value for each query selectivity factor;

e) matching any conditional selectivity expression with stored statistics to obtain statistics that can estimate the selectivity of the conditional selectivity expressions by:

i) compiling a set of candidate statistics that can be used to estimate the selectivity of the conditional selectivity expression;

ii) selecting candidate statistics to estimate the selectivity of the conditional selectivity expression based on a selection criteria; and

iii) using the statistics to obtain an estimated selectivity value; and

47. The method of claim 46 wherein the selection criteria for a candidate statistic is determined by computing a number of independence assumptions that are made when the candidate is used to estimate the selectivity of the conditional selectivity expression and the selection criteria is to select the candidate that results in the least number of independence assumptions.

48. The method of claim 46 wherein the selection criteria for a candidate statistic is determined by comparing the candidate statistic with a base statistic over the same column as the candidate statistic and assigning a difference value to the candidate statistic based on a level of difference between the candidate statistic and the base statistic.

49. The method of claim 46 wherein the step of compiling a set of candidate statistics is performed by including statistics that are on results of queries having the same tables referenced by the conditional selectivity expression or a subset of the tables referenced by the conditional selectivity expression and the same predicates over the tables referenced in the conditional selectivity expression or a subset of the predicates over the tables referenced in the conditional selectivity expressions.

50. The method of claim 46 wherein the product formed in step c) further comprises a query selectivity factor and wherein steps b) - f) are recursively performed to determine a selectivity value for the query selectivity factor in step c).

51. The method of claim 46 wherein steps b) - f) are recursively performed until a non-separable query selectivity that can only be decomposed into a single conditional selectivity expression results.

52. The method of claim 46 comprising the step of associating an error with the estimated selectivity value that is based on an accuracy with which the statistic matched with the conditional selectivity expression can estimate its selectivity.

53. The method of claim 52 comprising the step of combining the error associated with each conditional selectivity expression to obtain an estimated error for the selectivity estimation for the query.

54. A a computer readable medium having computer executable instructions stored thereon for approximating a number of tuples returned by a database query that comprises a set of predicates that each reference a set of database tables, the method comprising the steps of:

a) expressing the query as a query selectivity;

b) determining if the query is separable and if so separating the query selectivity by separating the predicates that reference different sets of database tables to form a product of query selectivity factors that reference different sets of database tables;

c) if the query is not separable, repeatedly decomposing the query selectivity to form a product that comprises a conditional selectivity expression to generate alternative products and wherein one of those products is selected to estimate the selectivity of the query;

d) recursively performing steps b) - f) to determine a selectivity value for each query selectivity factor;

e) matching any conditional selectivity expression with stored statistics to obtain statistics that can estimate the selectivity of the conditional selectivity expressions by:

i) compiling a set of candidate statistics that can be used to estimate the selectivity of the conditional selectivity expression;

ii) selecting candidate statistics to estimate the selectivity of the conditional selectivity expression based on a selection criteria; and

iii) using the statistics to obtain an estimated selectivity value; and

55. The computer readable medium of claim 55 wherein the selection criteria for a candidate statistic is determined by computing a number of independence assumptions that are made when the candidate is used to estimate the selectivity of the conditional selectivity expression and the selection criteria is to select the candidate that results in the least number of independence assumptions.

56. The computer readable medium of claim 55 wherein the selection criteria for a candidate statistic is determined by comparing the candidate statistic with a base statistic over the same column as the candidate statistic and assigning a difference value to the candidate statistic based on a level of difference between the candidate statistic and the base statistic.

57. The computer readable medium of claim 55 wherein the step of compiling a set of candidate statistics is performed by including statistics that are on results of queries having the same tables referenced by the conditional selectivity expression or a subset of the tables referenced by the conditional selectivity expression and the same predicates over the tables referenced in the conditional selectivity expression or a subset of the predicates over the tables referenced in the conditional selectivity expressions.

58. The computer readable medium of claim 55 wherein the product formed in step c) further comprises a query selectivity factor and wherein steps b) - f) are

recursively performed to determine a selectivity value for the query selectivity factor in step c).

59. The computer readable medium of claim 55 wherein steps b) - f) are recursively performed until a non-separable query selectivity that can only be decomposed into a single conditional selectivity expression results.

60. The computer readable medium of claim 55 comprising the step of associating an error with the estimated selectivity value that is based on an accuracy with which the statistic matched with the conditional selectivity expression can estimate its selectivity.

61. The computer readable medium of claim 60 comprising the step of combining the error associated with each conditional selectivity expression to obtain an estimated error for the selectivity estimation for the query.